Testing goodness-of-fit of random graph models

Villő Csiszár[1] Péter Hussami[2] János Komlós[3] Tamás F. Móri[1,5]
Lídia Rejtő[2,4] Gábor Tusnády[2]

submitted: May 4, 2012,
revised: November 8. 2012

Abstract

Random graphs are matrices with independent $0-1$ elements with probabilities determined by a small number of parameters. One of the oldest model is the Rasch model where the odds are ratios of positive numbers scaling the rows and columns. Later Persi Diaconis with his coworkers rediscovered the model for symmetric matrices and called the model beta. Here we give goodnes-of-fit tests for the model and extend the model to a version of the block model introduced by Holland, Laskey, and Leinhard.

1 Introduction

Let n be a positive integer, $1 \le i, j \le n$, and $\varepsilon(i,j)$ independent random variables such that $\varepsilon(i,j) = \varepsilon(j,i)$ and $\varepsilon(i,i) = 0$, furthermore

$$P(\varepsilon(i,j) = 1) = p_{i,j} = p + p_i + p_j, \quad 1 \le i < j \le n, \tag{1}$$

where the sum of the p_i-s is zero. The least square estimate $\hat{p}$ of p is the average of the epsilons, and the least square estimate of p_i is the average of the differences $\varepsilon(i,j) - \hat{p}$. The modification of the model for non-symmetric matrices is straightforward, and in that case the statistical inference is practically a two-way analysis of variance. Perhaps this is the simplest random graph model but it shares the inconvenient property of many other random graph models that it is hard to ensure that edge probabilities remain in the interval $(0,1)$. If we use the odds

$$r_{i,j} = \frac{p_{i,j}}{1 - p_{i,j}}, \tag{2}$$

[1]Eötvös Loránd University, Budapest, Hungary
[2]Alfréd Rényi Mathematical Institute of the Hungarian Academy of Sciences, Budapest,Hungary
[3]Rutgers University, Department of Mathematics, New Brunswick, New Jersey, USA
[4]University of Delaware, Statistics Program, FREC, CANR, Newark, Delaware, USA
[5]Tamás F. Móri's research was supported by OTKA grant 12574

instead of the probabilities, then it is enough to ensure the positivity of $r_{i,j}$-s. This is the case in the model introduced by George Rasch [31]. Historically the odds were defined as the ratios of scaling factors for rows and columns but we prefer the multiplicative form

$$r_{i,j} = \beta_i \gamma_j \tag{3}$$

for non-symmetric and

$$r_{i,j} = \beta_i \beta_j \tag{4}$$

for symmetric case. Statistical investigation of the model started with Andersen [1] (see also [21, 30, 33]) and later Persi Diaconis with his coworkers rediscovered the model and introduced the name *beta-model* for its parameter. The model has many attractive properties (see in [2, 4, 5, 6, 8, 28]):

– degree sequences are sufficient statistics

– the model covers practically all possible expected degree sequence

– the conditional distribution of the graphs on condition of a prescribed degree sequence is uniform on the set of all graphs with the given degree sequences.

Statistically inference emerged from Gaussian distribution and later was extended to random variables in Euclidean spaces but the statistical inference on discrete structures is rather sparse ([7, 15, 16, 19, 26]). Mathematical investigation of graphs has its own history. Nowadays instead of graphs we are speaking of networks ([27]) where the most investigated model is the stochastic block model introduced by Holland, Laskey, and Leinhard ([18]). Here the vertices are labeled by small numbers or colors and edge probabilities depend only on the labels ([3, 17]). With an eye on preferential attachment where degree sequences follow scale-free power-law the block model was criticized because it has moderated flexibility on degree sequences. Chung, Lu, and Vu [14] introduced a model with independent vertices, Chaughuri, Chung, and Tsiatas ([10]) introduced the *planted partition model* (see also [25]). Karrer and Newman [20] proposed and other extension of the block model. A natural extension of these models is the unification of the beta and block models:

$$r_{i,j} = b(i, c(j))b(j, c(i)), \tag{5}$$

where $b(.,.)$ is a positive matrix with n rows and k columns, and $c(i)$ is the label of the i-th vertex i.e. it is an integer between 1 and k. We call the model *k-beta model.* The estimation of the labels in block models is possible by the spectral method ([32]). It is generally believed that eigenvectors and eigenvalues of the matrix $\varepsilon(i, j)$ tells everything of the structure of the graph ([10, 12, 13, 22, 23, 24]), while there are many attempts to provide more flexible models ([9, 29]).

2 Goodness-of-fit

We can not test edge-independence on a single graph. While i.i.d. sample is common in statistical inference, in case of graphs the sample generally means a copy of a graph. Perhaps the number one question in statistical inference is the following. Let

$$p_1, \dots, p_n \tag{6}$$

be an arbitrary given sequence of probabilities, and

$$\varepsilon_1, \dots, \varepsilon_n \tag{7}$$

be independent $0-1$ variables such that $P(\varepsilon_i = 1) = p_i$. Can we test the model? A randomized answer is the following. Let

$$u_1, \dots, u_n \tag{8}$$

independent and uniformly distributed in $(0, 1)$. Then

$$x_i = p_i u_i \varepsilon_i + (1 - \varepsilon_i)(p_i + (1 - p_i)u_i), \quad i = 1, \dots, n \tag{9}$$

are independent and uniformly distributed in $(0, 1)$, what we can test. An other, more practical solution is ordering the the pairs (p_i, ε_i) according to the p_i-s in increasing order and compare their partial sums. Or we can clump them into blocks of small number and compare again the sums. All these possibilities hold for graphs with estimated edge probabilities. Let us partition the edges of the complete graph according to the blocks formed with respect to the edge probabilities. In each portion the edge probabilities are close to each other whence the $\varepsilon_{i,j}$-s corresponding to that portion behave like a pure random graph. what we again can test e.g. by their sums on subsets of vertices.

Blitzstein and Diaconis ([6, 11]) propose for testing the beta model the following general procedure. Let us choose any graph statistic and determine it on our graph. Let us generate as many graph we can with the same degree sequence as the investigated graph has according to the uniform distribution, and let us calculate the chosen statistics. If the value of the sample graph is inside the generated numbers, we accept the beta model, otherwise reject it. One can ask, are there any effect of the choose on the power of the test?

We have found by computer simulations that graphs generated by beta model have only one eigenvalue proportional with n, all the others are of order $\sqrt{n}$. We think that it is a characteristic property of beta graphs. One wonders that

– if beta model covers all possible degree sequences

– the conditional distribution is uniform over graphs sharing the same degree sequence,

then how is possible that graph behaves differently from typical graphs generated by beta model? Of course there are graphs having many large eigenvalues. But where are they coming

from once beta model can generate all the graphs? A possible solution of the catch is the following.

Let us generate a meta graph from graphs sharing the same degree sequence. Let us say that neighborhood in this meta graph is given by on single swap. If we have four vertices A, B, C, D in a graph such that AC, BD is and edge but AD, BC is not, then changing existence into non existence among these edges we form a new graph with the same degree sequence. The degree of a graph in this meta graph goes parallel with the second largest eigenvalue: typical beta model graphs have minimal degree and any increase in their degree results in a more complicated eigenvalue structure. Perhaps the degree in the meta graph is the most characteristic statistic for beta model.

3 The k-beta model

The maximum likelihood equations for the parameters $b(.,.)$ in (5) say that the expected values of degrees *inside* all the subgraph with a given pair of labels should be the same us in the given graph. This is the case when the labels are known. With unknown labels we can form a two-level optimization: for each label set first to determine the parameters $b(.,.)$ next changing a small number of labels and repeat the calculation of the parameters. But the procedure is slow even for graphs of moderate sizes. Spectral methods available for block models fail for coloring k-beta models because the model lose the well pronounced checkerboard character of block models. It is the ANOVA what offers an applicable algorithm. For any set C of labels $c(.)$ let us calculate the statistic

$$Q(C) = \sum_{i=2}^{n} \sum_{j=1}^{i-1} (\varepsilon(i,j) - u(c(i),c(j)) - v(i,c(j)) - v(j,c(i)))^2, \tag{10}$$

where

$$u(s,t) = \frac{\sum_{c(i)=s} \sum_{c(j)=t} \varepsilon(i,j)}{\sum_{c(i)=s} \sum_{c(j)=t} 1}, \tag{11}$$

and

$$v(i,t) = \frac{\sum_{c(j)=t} (\varepsilon(i,j) - u((c(i),t))}{\sum_{c(j)=t} 1}. \tag{12}$$

$Q(C)$ is the sum of two way ANOVA sum of squares calculated independently for subgraphs defined for pairs of labels. Starting from a uniform random set C of labels on the vertices and perturbing small number of labels in the individual steps a simple greedy optimization results in a good set of labels, which is close to the original (true) labels.

For evaluating the character of a random graph we use the number

$$\exp(-\frac{\sum_{i=2}^{n} \sum_{j=1}^{i-1} (p(i,j) \log p(i,j) + (1-p(i,j)) \log(1-p(i,j))}{n(n-1)/2}) \tag{13}$$

We call it *delogarithmed average entropy* or DAE. This is a number between 1 and 2. If it is close to one the graph is almost deterministic: the probabilities are close to 0 or 1. In checkerboard block models it means that empty and full subgraphs are amalgamated together. If DAE is close to 2 then the graph has no structure at al. DAE depends on edge density, too. The above tendency is valid for edge density $\frac{1}{2}$, for other edge densities the cut point is closer to 1. According to our experience if DAE is smaller then 1.9 while edge density is half, then we are able to reconstruct the original labels. For these graphs the number of non-trivial eigenvalues is $2k-1$, thus the spectrum determines the number of different labels.

The k-beta model has a sister model

$$r_{i,j} = \sum_{s=1}^{k} b(i,s)b(j,s) \tag{14}$$

what we call *small odds rank* model. Strictly speaking we ought to redefine the diagonal of odds matrix, but perhaps the name is permissible without doing so. The maximum likelihood estimation of parameters in small odds rank models is straightforward and the block structure is detectable in the estimated parameters. Actually the block model is in the intersection of k-beta and small odds rank models, thus if there is any block structure in the graph it is detectable even in fitting k-beta model to the graph. But if there is no block structure and we are trying to use ANOVA coloring for a small odds rank graph then the algorithm is no longer stable, it results in different local minima in each runs.

References

[1] E. B. Andersen, Sufficient statistics and latent trait models, *Psychomretrika* **42** 1 (1977), 69–81.

[2] P. J. Bickel and A. Chen, A nonparametric view of network models and Newman-Girvan and other modularities, *Proceedings of the National Academy of Sciences* **106** , 50 (2009), 21068–21073.

[3] P. Bickel, D. Choi, X. Chang, and H. Zhang, Asymptotic normality of maximum likelihood and its variational approximation for stochastic blockmodels, Preprint available at arXiv:1207.0865v1 [math.ST] 4 Jul 2012, Submitted to the *Annals of Statistics*

[4] A. Barvinok and J. A. Hartigan, An asymptotic formula for the number of non-negative integer matrices with prescribed row and column sums, Preprint, available at http://arxiv.org/abs/0910.2477, (2009).

[5] A. Barvinok and J.A. Hartigan, The number of graphs and a random graph with a given degree sequence, Preprint, available at http://arxiv.org/abs/1003.0356, (2010).

[6] J. Blitzstein and P. Diaconis, A sequential importance sampling algorithm for generating random graphs with prescribed degrees, *Journal of Internet Mathematics*, **6** (2010), 489–522.

[7] M. Bolla and G. Tusnády, Spectra and optimal partitions of weighted graphs, *Discrete Mathematics* **128**, (1994), 1–20.

[8] S. Chatterjee, P. Diaconis and A. Sly, Random graphs with a given degree sequence, Preprint, available at arXiv: 1005.1136v3 [math.PR], (2010), submitted to the Annals of Statistics.

[9] S. Chatterjee and P. Diaconis, Estimating and understanding exponential random graph models, Preprint available at arXiv:1102.2650v3[math.PR] 6 Apr 2011.

[10] K. Chaudhuri, F. Chung, and A. Tsiatas, Spectral clustering of graphs with general degrees in the extended planted partition model, *Journal of Machine Learning Research* **1** (2012) 1–23.

[11] Y. Chen, P. Diaconis, S. P. Holmes, and J. S. Liu, Sequential Monte Carlo methods for statistical analysis of tables, *Journal of the American Statistical Association* **100:469** (2005) 109–120.

[12] F. Chung, Spectral gragh theory, *American Mathematical Society, Providence, RI* 1997.

[13] F. Chung and L. Lu, Complex graphs and networks, *American Mathematical Society, Boston, Massachusetts,* 2006.

[14] F. Chung, L. Lu, and V. Vu, Spectra of random graphs with given expected degrees, *Proceedings of the National Academy of U.S.A* **27** (2003), 6313–6318.

[15] V. Csiszár, L. Rejtő and G. Tusnády, Statistical inference on random structures, *Horizon of Combinatorics* , (eds. Győri,E. et al.), (2008), 37–67.

[16] V. Csiszár, P. Hussami, J. Komlós, T. Móri, L. Rejtő, and G. Tusnády, When the degree sequence is a sufficient statistic, *Acta Mathematica Hungarica* **134** (2011) 45–53.

[17] C. J. Flynn and P. O. Perry, Consistent biclustering, Preprint available at arXiv:1206.6927v1 [stat.ME] 29 Jun 2012.

[18] P. Holland, K. B. Laskey, and S. Leinhardt, Stochastic blockmodels: some first steps, *Journal of the American Statistical Association,* **76(373)** (1981), 33–50.

[19] P. Hussami, Statistical inference on random graphs, PhD thesis, (Central European University, Budapest, Hungary) (2010).

[20] B. Karrer and M. E. J. Newman, Stochastic blockmodels and community structure in networks, *Phys. Rev. E* **83** (2011), 016107

[21] J. M. Linacre, Predicting responses from Rasch measures, *Journal of Applied Measurement* **11** (2010).

[22] L. Lovász, Very large graphs, *Current Developments of Mathematica* 2008, 67-128.

[23] L. Lu and X. Peng, Spectra of edge-independent random graphs, Preprint available at arXiv:1204.6207v1 [math.CO] 27 Apr 2012.

[24] R. R. Nadakuditi and M. E. J. Newman, Spectra of random graphs with arbitrary expected degrees, Preprint available at arXiv:1208.1275v1 [cs.SI] 6 Aug 2012.

[25] E. Mossel, J. Neeman, and A. Sly, Reconstruction and estimation in the planted partition model, Preprint, available at arXiv: 1202.1499v4 [math.PR] 22 Aug 2012.

[26] T. Nepusz, L. Négyessy, G. Tusnády, and F. Bazsó, *Dynamic System and its Applications* **18** (2009) 335–362.

[27] M. Newman, A.-L. Barabási and D. Watts, *The structure and dynamics of networks, (Princeton studies in complexity)*, Princeton University Press, 2007.

[28] M. Ogawa, H. Hara, and A. Takemura, Graver basis for an undirected graph and its application to testing the beta model of random graphs, *Ann. Ist. Stat. Mat.* 2012.

[29] G. Palla, L. Lovász and T. Vicsek, Multifractal network generator, *Proceedings of the National Academy of Sciences* **107**, 17 (2010), 7641–7645.

[30] I. Ponicny, Nonparametric goodness-of-fit tests for the Rasch model, *Psychometrika* **66** (2001), 437–460.

[31] G. Rasch, Probabilistic models for some intelligence and attainment tests, *Coppenhagen: Danmarks Paedogiske Institut,* 1960.

[32] K. Rohe, S. Chatterjee, and B. Yu, Spectral clustrering and high-dimensional stochastic block model, *Annals of Statistics* **39** (2011) 1878–1915.

[33] N. Verhelst, Testing the unidimensionality assumption of the Rasch mode, *Measurement and Research Department Reports* 2002.